Research Report

A Modified *a*-Stratified Method for Computerized Adaptive Testing

ETS RR–19-10

Lixiong Gu
Guangming Ling
Yanxuan Qu

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Modified *a*-Stratified Method for Computerized Adaptive Testing

Lixiong Gu, Guangming Ling, & Yanxuan Qu

Educational Testing Service, Princeton, NJ

Research has found that the *a*-stratified item selection strategy (STR) for computerized adaptive tests (CATs) may lead to insufficient use of high *a* items at later stages of the tests and thus to reduced measurement precision. A refined approach, unequal item selection across strata (USTR), effectively improves test precision over the STR by allowing more items to be selected from the strata with higher *a*-parameter values. However, both approaches ignore the contribution of items' *c*-parameters to the information. This study proposes another procedure — maximum information STR (MISTR) — that groups items based on the maximum amount of Fisher information an item can provide. This information is a function of its *a*- and *c*-parameters. MISTR can be further modified to select more items from strata with high *a*-parameter values (unequal MISTR [UMISTR]). This study evaluated and compared MISTR, UMISTR, STR, and USTR on two aspects of the CAT performance: (a) quality of θ estimation and (b) effectiveness in item pool usage. The results showed that both the MISTR and UMISTR approaches produced more precise ability estimation than the STR approach when the test length was longer and when an item-exposure-control procedure was used. The UMISTR produced slightly less precise ability estimation than USTR but led to fewer underused items, indicating a more balanced use of the item pool. These findings suggest that MISTR and UMISTR can be viable alternatives to STR and USTR.

Keywords Computerized adaptive tests; *a*-stratified strategy; maximum information STR

doi:10.1002/ets2.12246

The *a*-stratified item selection strategy (STR; Chang & Ying, 1999) is among the most commonly used item selection methods in computerized adaptive tests (CATs). It partitions the item pool into strata according to an item's *a*-values and selects items from strata with low *a*-values in the beginning stages and items from strata with high *a*-values in the later stages. Within the same stratum, instead of selecting an item that maximizes Fisher information at ability estimate $\widehat{\theta}$, an item with a *b*-parameter closest to the $\widehat{\theta}$ is selected (Chang & Ying, 1999). This simpler criterion is used because Chang and Ying (1999) assumed that within a stratum, *a*-parameter values of the items are similar, thus matching *b* with $\widehat{\theta}$ closely approximates maximizing item information. Compared to the traditional item selection procedures that are based on maximum Fisher information, the STR procedure achieves a more balanced item usage and still maintains acceptable test precision (Deng, Ansley, & Chang, 2010).

Research has indicated four prominent shortcomings of the STR procedure. First, items in the strata with high *a*-values tend to have high *b*-values (more difficult). A shortage of lower *b* items in those strata may cause low *b* items to be selected more frequently and lead to less accurate ability estimates (Chang, Qian, & Ying, 2000; Parshall, Davey, & Nering, 1998). Second, the STR procedure tends to underuse highly discriminating (high *a*) items in later stages and thus sacrifices test precision. A refined stratified procedure, unequal item selection across strata (USTR; Deng & Chang, 2001), was developed to allow more items to be selected from the high *a* strata and fewer items to be selected from the low *a* strata. Third, because strata are typically constructed by grouping an equal number of items into each stratum, the range of *a*-parameter values for items in the same stratum could be fairly wide. This, to some extent, defeats the idea of the STR procedure using low *a*-parameter items in the early stages and high *a*-parameter items in the later stages of the CAT administration. Fourth, with the three-parameter logistic (3PL) item response theory model, item information at an ability level that matches the *b*-parameter is determined by both the *a*- and *c*-parameters. Thus considering both *a*- and *c*-parameters when grouping items into strata seems more likely to place items that provide similar information in the same stratum.

To address these issues for STR, this study proposes a modified approach maximum information STR (MISTR), which groups items into strata based on the maximum amount of information (MI_i) an item can provide (Lord, 1980). As

Corresponding author: L. Gu, E-mail: lgu@ets.org

depicted in Equation 1,

$$\mathrm{MI}_i = \frac{D^2 a_i^2}{8\left(1-c_i\right)^2}\left[1 - 20c_i - 8c_i^2 + \left(1+8c_i\right)^{3/2}\right]. \quad (1)$$

MI_i is a function of both the *a*- and *c*-parameters; therefore the MISTR approach takes into account both parameters when stratifying the pool rather than simply grouping an equal number of items into each stratum. One can choose to stratify the item pool based on any statistic that is a function of both *a*- and *c*-parameters, but MI_i is chosen here because the calculation formula is readily available and shows the maximum amount of information an item can provide.

The following is one example of how MISTR groups items into strata: Items with $\mathrm{MI}_i < .2$ are grouped into Stratum 1, those with MI_i between .2 and .4 are grouped into Stratum 2, those with MI_i between .4 and .6 are grouped into Stratum 3, and those with $\mathrm{MI}_i > .6$ are grouped into Stratum 4. This grouping approach leads to having items that provide a similar level of measurement precision in each stratum, although the number of items varies across strata. For each test, the number of items selected from a particular stratum is proportional to the ratio of the number of items in that stratum (stratum size) over the number of items in the whole pool. Alternatively, like the USTR approach Deng and Chang (2001) proposed, more items can be selected from strata with higher maximum information to further improve the measurement precision (i.e., UMISTR).

By grouping items that provide similar maximum information into the same stratum, it is expected that MISTR and UMISTR item selection methods follow more closely the initial idea of Chang and Ying (1999), which uses low discriminating items early in the test and saves items with high discrimination for use later in the test. It, therefore, may achieve a more effective item pool usage with similar quality in ability estimation. By comparing the performance of MISTR and UMISTR procedures with that of other procedures (STR and USTR) using simulated CAT administrations under various practical conditions—including constraints such as item-exposure control—this study attempts to answer the following research question: Do MISTR and UMISTR produce more accurate θ estimates and more effective item pool usage in comparison to STR and USTR methods?

Method

Data

A pool of 500 items was simulated based on the item characteristics of an operational item pool for a large-scale CAT. Specifically, the items' *a*-parameters were drawn from a log-normal distribution with a mean of 0.8 and a standard deviation of 0.3. The *b*-parameters were drawn from a normal distribution with a mean of −0.4 and a standard deviation of 1.05, and the *c*-parameters were drawn from a beta distribution with a mean of 0.18 and a standard deviation of 0.09. The *a*- and *b*-parameters were simulated to have a Pearson correlation coefficient of .25 ($\rho_{ab} = .25$).

Examinee abilities were randomly sampled from a normal distribution, $N(0,1)$. The examinee responses were generated according to the 3PL model.

Design

Table 1 illustrates the variables manipulated in this simulation study. Test conditions resulted from combinations of two test lengths (40 and 60 items), one practical constraint (exposure control), and four item-selection procedures (STR, USTR, MISTR, and UMISTR). Two conditions of fixed test length (40 and 60 items) were used to investigate the performance of different item selection procedures in conditions where more or less freedom of item selection was allowed. The Sympson–Hetter (SH) exposure-control procedure (Sympson & Hetter, 1985) was used to evaluate how each procedure behaved under conditions with or without item-exposure control.

The SH exposure-control procedure uses a conditional selection process to control item exposure. It assigns each item an exposure-control parameter value that ranges from 0 to 1 and is predetermined based on the frequency of item selections in an iterative CAT simulation. Items that are more likely to be selected are assigned smaller exposure-control parameter values, while items less likely to be selected are assigned larger values. During the test operations, the exposure-control parameter of the selected item is compared to a random number drawn from a uniform distribution ranging from 0 to1. If the exposure-control parameter is greater than the random number, the item is administered. If it is smaller, the item is put back into the item pool and the same process is applied to the next best item. The item exposure-control

Table 1 Specifications of the Computerized Adaptive Test Simulation Design

Variable	Specification
Test length	40, 60
Item selection	STR, USTR, MISTR, UMISTR
Examinee distribution	$N(0,1)$, $N = 3{,}000$ examinees
Exposure control	No exposure control Sympson – Hetter (with the target exposure rate of 0.2 or less)
Content balancing	No content balancing
Number of strata	4

Note. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

parameter is like a threshold. By controlling the thresholds, the SH method limits the administration of frequently used items in CAT and ensures that infrequently used items are used more frequently.

Under each of the four stratification conditions, the item pool was divided into four strata. The same stratification method was applied for STR and USTR, where the item pool (500 items) was evenly divided into four strata based on an item's *a*-parameters. As a result, each stratum had 125 items. A slightly different stratification method was used for both MISTR and UMISTR, where the item pool was divided based on an individual item's maximum information MI_i. More specifically, items with an MI_i less than .2 were grouped into Stratum 1, those with an MI_i between .2 and .4 were grouped into Stratum 2, and those with an MI_i between .4 and .6 were grouped into Stratum 3. All items with an MI_i above .6 were grouped into Stratum 4. As a result, this method put 132 items in Stratum 1, 197 items in Stratum 2, 106 items in Stratum 3, and 65 items in Stratum 4.

Under the STR conditions, 10 items were selected from each stratum for the 40-item tests and 15 items were selected for the 60-item tests. Under the USTR conditions, the number of items selected from Strata 1 – 4 were 6, 8, 12, and 14 for 40-item tests and 9, 12, 18, and 21 for 60-item tests. This approach was used by Deng et al. (2010). Under MISTR and UMISTR conditions, the number of items in a test selected from each stratum was determined proportionally to the total number of items in the stratum. The only difference between MISTR and UMISTR is how the unrounded numbers are treated. With MISTR, the number of items is rounded for Strata 2 – 4, and the rest of the items are selected from Stratum 1. This forces the CAT to use more low *a*-parameter items in the early stages of the test. With UMISTR, the number of items to be selected from each stratum is truncated to the nearest integer for Strata 1 – 3, and the rest of the items are selected from Stratum 4. This allows the CAT to use more high *a*-parameter items in the later stages of the test.

USTR draws 65% of the items (26 items for a 40-item test) in each CAT from the two strata with higher *a*-values (those two strata contain 50% of the items in the pool). UMISTR, in contrast, draws 37.5% of the items (15 items for a 40-item test) from the two strata with high *a*-parameter values (34% of the items in the pool). This indicates that UMISTR uses only slightly more items from strata where highly discriminating items are grouped when compared to MISTR, but USTR uses many more highly discriminating items than STR (Table 2).

Table 3 presents the descriptive statistics of the item parameters in each stratum by item selection method. As shown in the table, the MISTR and UMISTR procedures have strata with greater mean *a*- and *b*-parameter values than the STR and USTR procedures and with similar mean *c*-parameter values as the STR and USTR procedures. With MISTR and UMISTR procedures, the standard deviation of the *a*-parameter values for each stratum is also higher than those standard deviations in corresponding strata with the STR and USTR procedure. This indicates that items in strata grouped by maximum information have a wider range of *a*-parameters.

Each examinee was assigned an initial ability estimate of −0.5 under all stratified conditions (STR, USTR, MISTR, and UMISTR). Deng et al. (2010) used the same strategy in their study. The first item was randomly selected from among the 10 items in the first stratum with *b*-values most closely matching the θ-value −0.5. This item selection strategy is intended to eliminate similar item sequences across examinees early in the test. After the first item administration, the ability estimate was updated, and the subsequent items were selected — one at a time — by matching *b*- and θ-values (i.e., selecting an item that minimizes the absolute difference of its *b*-parameter and θ estimates).

Table 2 Number of Items in Item Pool for Each Stratum and Number of Items in the Test Selected From Each Stratum

	Stratum				
Item Selection Method	1	2	3	4	Total
STR					
Items in pool	125	125	125	125	500
Items in test (STR)	10	10	10	10	40
Items in test (USTR)	6	8	12	14	40
Items in test (STR)	15	15	15	15	60
Items in test (USTR)	9	12	18	21	60
MISTR					
Items in pool	132	197	106	65	500
Items in test (MISTR)	11	16	8	5	40
Items in test (UMISTR)	10	15	8	7	40
Items in test (MISTR)	15	24	13	8	60
Items in test (UMISTR)	15	23	12	10	60

Note. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

Table 3 Descriptive Statistics of Item Parameters by Stratum

	Stratum				
Item Parameter	1	2	3	4	Total
STR/USTR, *M* (*SD*)					
a	0.526 (0.072)	0.697 (0.036)	0.862 (0.057)	1.127 (0.165)	0.803 (0.242)
b	−0.729 (1.018)	−0.669 (1.119)	−0.394 (1.063)	0.005 (1.002)	−0.447 (1.088)
c	0.193 (0.114)	0.179 (0.091)	0.168 (0.089)	0.201 (0.107)	0.185 (0.101)
Items in pool	125	125	125	125	500
MISTR/UMISTR, *M* (*SD*)					
a	0.545 (0.099)	0.755 (0.103)	0.957 (0.081)	1.221 (0.178)	0.803 (0.242)
b	−0.660 (1.039)	−0.655 (1.117)	−0.146 (0.902)	0.125 (1.072)	−0.447 (1.088)
c	0.228 (0.124)	0.182 (0.094)	0.154 (0.075)	0.160 (0.080)	0.185 (0.101)
Items in pool	132	197	106	65	500

Note. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

For each of the eight conditions with SH exposure control, a unique set of exposure-control parameters was derived through seven iterative simulations using a sample of 1,000 normally distributed abilities drawn from $N(0,1)$. The results from the final round of the simulation were taken as the exposure-control parameters for the simulated CAT administrations.

Evaluation Criteria

The performance of the item selection methods was evaluated on two aspects: quality of θ estimation and effectiveness in item pool usage. Those criteria were similar to the ones used by Chang and Ying (1999) and Reckase and He (2005). The overall quality of ability estimation was evaluated by examining the correlation between true and estimated θ-values, bias, and root mean square error (RMSE) of θ estimates. Effectiveness of the item pool usage was evaluated by observed item exposure rates (see the plots in Appendix A), skewness of the item exposure rate distribution, the number of over- and underexposed items, and test overlap rate.

Correlation Between True and Estimated θ Values ($\rho_{\theta\hat{\theta}}$)

The correlation coefficient between the true and estimated θ-values can be interpreted as the correlation associated with the observed and true scores on the test (Lord, 1980, p. 52).

Bias and Root Mean Square Error

These quantities are defined as follows:

$$\text{Bias} = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{\theta}_j - \theta_j\right), \tag{2}$$

and

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{\theta}_j - \theta_j\right)^2}, \tag{3}$$

where N is the number of simulees and $\widehat{\theta}_j$ is the ability estimate of the jth simulee that has true ability level θ_j.

To evaluate the test precision along the continuum of the θ scale, simulees were first grouped together if their rounded true θ-values were the same to the 10th decimal place. The following three statistics were calculated for each group of simulees: box plots of the ability estimates (see the plots in Appendix B), average test information (see the plots in Appendix C), and the conditional standard error of measurement (CSEM; see the plots in Appendix D).

Conditional Standard Error of Measurement

The CSEM is calculated by the formula

$$\text{CSEM} = \sqrt{\frac{1}{N_g}\sum_{j=1}^{N_g}\left(\widehat{\theta}_{g_j} - \overline{\theta}_g\right)^2}, \tag{4}$$

where N_g is the number of adaptive tests administered within θ group g and

$$\overline{\theta}_g = \frac{1}{N_g}\sum_{j=1}^{N_g}\widehat{\theta}_{g_j}$$

is the average ability estimates over the N_g CATs within θ group g.

Skewness of Item-Exposure Rate Distribution

A χ^2-like statistic proposed by Chang and Ying (1999) was used to capture the discrepancy between the observed and the ideal item exposure rates. This statistic can be considered a measure of skewness for the item-exposure rate distribution and an indication of the efficiency of item pool usage. It is defined as follows:

$$\chi^2 = \sum_{i=1}^{n}\frac{\left(r_i - L/n\right)^2}{L/n}, \tag{5}$$

where r_i is the observed exposure rate for the ith item, L is the test length, and n is the number of items in the item pool. L/n is the average exposure rate of the item pool and the desirable uniform rate for all items. A low χ^2 value implies that most of the items are fully used.

Number of Overexposed Items

A moderate level of item exposure rate is generally desired for all the items in an item pool to maintain security of the items and validity of the test. A high exposure rate for an item means an increased risk of the item being known by prospective examinees. In this study, an item with an exposure rate greater than .2 is considered overexposed.

Number of Underexposed Items

A low item exposure rate means that the item is rarely used. An item pool with too many items with too low an exposure rate will likely have many overexposed items, which is a sign of the unbalanced usage and underutilization of some of the items in the pool. In this study, an item with an exposure rate lower than .02 is considered underexposed.

Test Overlap Rate

Test overlap rate is the expected number of common items encountered by two randomly selected examinees divided by the expected test length. Ideally, the number of common items between any two randomly sampled examinees should be minimized. Equation 6 summarizes the calculation of test overlap rate (Chen, Ankenmann, & Spray, 1999):

$$\overline{T} = \frac{\sum_{i=1}^{n} \binom{m_i}{2}}{L\binom{N}{2}} = \frac{\sum_{i=1}^{n} m_i\left(m_i - 1\right)}{LN\left(N-1\right)}, \tag{6}$$

where N denotes the number of simulees and is therefore the number of fixed-length CATs administered, L is the test length, n is the total number of items in the pool, and m_i is the frequency that item i is administered across all N CATs.

Results

The results are organized by condition of practical constraints: with or without exposure control. Performances of the four procedures are summarized in terms of relative quality of ability estimates and item pool usage.

Without Exposure Control

Table 4 summarizes the evaluation criteria for the performance of STR, MISTR, USTR, and UMISTR under 40-item and 60-item conditions when exposure control is not imposed on the item selection procedure. Figures A1 and A2 show scatterplots of item parameters in the pool against items' exposure rates for the same conditions. Figures B1 and B2 show box plots of ability estimation conditional on true ability levels. Figures C1 and C2 present the average test information function conditional on true ability levels, and Figures D1 and D2 plot the conditional standard errors of measurement.

It can be seen from Table 4 that, under 40-item conditions, STR and MISTR perform similarly well, with STR being slightly better on all the evaluation criteria. USTR and UMISTR both obtain smaller RMSE than STR and MISTR, indicating an improved quality for ability estimation for both methods by using more high *a*-parameter items at the later stages of the test. In general, UMISTR has slightly greater bias, greater RMSE, and a lower correlation between the true and estimated ability than USTR. This, however, may be due to the fact that UMISTR is more conservative in using high *a*-parameter items than USTR, which results in a higher number of underused items (i.e., items with exposure rate lower than .02). The scatter plots show that MISTR and UMISTR use more items within the ability range of 0.8 and 1.5. Figure C1 also indicates that average test information differs the most within the ability range from −1.0 to 2.0. USTR observed the highest average test information at all ability levels, followed by UMISTR, The average test information produced by MISTR and STR was similar.

The 60-item conditions show patterns similar to the 40-item conditions. USTR produces the least RMSE for ability estimates, which is highly desirable, although it still leads to the largest number of underused items and seems to have the highest test overlap rate among the four. UMISTR produces slightly better quality on ability estimates than MISTR and STR, which produces the largest RMSE for ability estimates.

With Sympson–Hetter Exposure Control

Table 5 lists the summary of the evaluation criteria for the performance of STR, MISTR, USTR, and UMISTR under 40-item and 60-item conditions when SH exposure control is imposed on the item selection procedure. Figures A3 and A4 show the scatter plots of item parameters in the pool against items' exposure rates for the same conditions. Figures B3 and B4 show the box plot of ability estimation conditional on true ability levels. Figures C3 and C4 present the average test information function conditional on true ability levels, and Figures D3 and D4 plot the CSEMs.

Table 5 shows that SH is effective in controlling the item-exposure rates, resulting in a much smaller number of overexposed items, although it seems to also increase the number of items that are underexposed. In general, the results are similar to the conditions where no exposure control is imposed. The STR and MISTR methods perform similarly. When SH exposure control is used, however, MISTR slightly outperforms STR by having smaller RMSE values, under both the

Table 4 Performance Summary for Item Selection Across Strata Versus Maximum Information Item Selection Across Strata Without Exposure Control

Item Selection Method	Bias	RMSE	χ^2	$N_{er<.02}$	$N_{er>.2}$	Overlap rate	$\rho_{\theta\hat{\theta}}$
40 Items							
STR	0.015	0.291	14.68	24	17	0.109	0.960
MISTR	0.016	0.292	16.44	31	21	0.113	0.959
USTR	0.021	0.277	15.84	68	9	0.111	0.964
UMISTR	0.025	0.281	16.33	38	20	0.112	0.962
60 items							
STR	0.005	0.231	16.59	6	55	0.153	0.974
MISTR	0.011	0.228	18.12	10	59	0.156	0.975
USTR	0.008	0.216	20.43	25	59	0.161	0.977
UMISTR	0.017	0.227	18.32	9	55	0.156	0.975

Note. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR; RMSE = root mean square error.

Table 5 Performance Summary for Item Selection Across Strata Versus Maximum Information Item Selection Across Strata With Sympson – Hetter Exposure Control

Item Selection Method	Bias	RMSE	χ^2	$N_{er<.02}$	$N_{er>.2}$	Overlap rate	$\rho_{\theta\hat{\theta}}$
40 items							
STR-SH	0.015	0.293	15.40	47	4	0.110	0.959
MISTR-SH	0.009	0.291	16.74	47	4	0.113	0.959
USTR-SH	0.011	0.268	18.44	88	0	0.117	0.966
UMISTR-SH	0.020	0.277	17.08	54	6	0.114	0.963
60 items							
STR-SH	0.015	0.238	13.33	13	46	0.146	0.973
MISTR-SH	0.008	0.236	14.61	20	35	0.149	0.973
USTR-SH	0.011	0.221	17.18	39	18	0.154	0.976
UMISTR-SH	0.010	0.230	14.64	15	24	0.149	0.974

Note. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR; SH = Sympson – Hetter; RMSE = root mean square error.

conditions with 40 items and 60 items. On the other hand, STR has smaller test overlap rates and smaller χ^2 values in both conditions, indicating a slightly more balanced item usage. USTR and UMISTR both lead to smaller RMSE and a higher correlation between true and estimated θs than STR or MISTR. The trade-off is a slightly unbalanced item pool use and a higher test overlap rate. UMISTR seems to produce a well-balanced performance, with slightly larger χ^2 values and test overlap rate, similar number of overexposed and underexposed items, and an improved ability estimate quality (RMSE) than the STR and MISTR methods. In comparison to USTR, it has slightly higher RMSE values, but a more balanced item pool usage and less underexposed items.

Discussion

The *a*-stratified method (Chang & Ying, 1999) is a simple but efficient item selection procedure to ensure that items with higher discriminating power are administered at the later stages of a CAT, when the ability estimates are more stable. Since it was proposed, there have been many modifications to address some issues it faced, such as overuse of low *b*-value items and underuse of the highly discriminating items. This study proposes another modification through the use of both *a*- and *c*-parameters to stratify the item pool in the hope of optimizing item pool usage and improving measurement precision. The idea is that by grouping items with similar maximum information together when their *b*-values match with ability estimates, the item would provide measurement precision that is intended in that stratum. In addition, when an item with a closer match in *b*-values has already been used, another item that has a slightly smaller *b*-value would provide a similar level of item information. Therefore, it is expected that this approach will lead to more precise estimation of ability levels. The results, however, seem to be mixed in this regard. In all simulation conditions, the STR

approach always produced the lowest χ^2 values, indicating a more balanced usage of the items in the pool. In three of the four conditions, however, the MISTR approach produced slightly more precise ability estimates (60-item tests without exposure control, 40-item and 60-item tests with exposure control). The UMISTR, by selecting more items at the strata with high *a*-parameters, performed better than STR or MISTR but slightly worse than USTR. This seems reasonable because UMISTR selects significantly more items from the strata with higher *a*-parameters. A future study may be needed to explore what would be the optimal number of items to select from each stratum under MISTR or UMISTR conditions.

One of the rationales behind using predetermined maximum information values to stratify an item pool is the concept of *bin* (Reckase & He, 2005). Bin is used to describe a boundary for certain characteristics of the items, so that items in the same bin are treated as interchangeable in test administration. For example, bins could be defined by items' *a*- and *b*-parameters so that items having *a*-parameters between 1.0 and 1.2 and *b*-parameters between 0.0 and 0.2 are considered to be in the same bin and to have similar psychometric properties and, therefore, can be administered interchangeably. The bins can also be defined by the content standards that items measure. The *a*-stratified approach, when items are selected by matching *b*-values, could be considered as an example of the bin concept. The MISTR goes a step further by taking into account the value of *c*-parameters when items are grouped, creating bins that consist of statistically more interchangeable items. A more refined MISTR approach may adopt the idea behind the *a*-stratified and *b*-blocking approach, which stratifies the item pool differently in different blocks of items that are first grouped by *b*-parameters. This approach may lead to stratification that makes more items with a low *b*-parameter available in strata with high *a*-parameter items. Future research could look into the use of different stratification rules for different blocks of *b*-parameters.

Another motivation for using maximum information to stratify an item pool is the hope of making optimal pool design feasible under an item pool stratification framework. Under the STR approach, the item pool is divided into strata with equal numbers of items. This method of stratification does not depend on item characteristics. In contrast, because the MISTR approach groups items providing similar item information together, distribution of the items in each stratum acts like a statistical blueprint for the item pool. Simulation studies may be conducted to determine the appropriate number of items in each stratum.

One concern about the effectiveness of this approach is tied to an issue with most CAT simulation studies in which the pool stratification and item selection are based on the true item parameter values (i.e., those used in generating item responses). It is well known that the *c*-parameter is often poorly estimated (Hambleton, Swaminathan, & Rogers, 1991) with the 3PL model. Even with 1PL or 2PL models, item parameters are estimated with error. It may be the case that any advantage achieved through consideration of an additional item parameter is canceled by estimation problems. To see how the four item-selection methods compare in practice, it may be desirable to use item parameter estimates based on realistic sample sizes. On the other hand, even if a 2PL model is used to model the items in the CAT pool, it is still helpful to consider the idea of grouping items based on the specific ranges of item parameters instead of putting an equal number of items in a stratum because of the potential advantages mentioned earlier, such as making items in a stratum more interchangeable and facilitating optimal item pool design.

This study has some limitations that could be improved upon in the future. First, only one CAT simulation was run for each condition, which may lead to limited generalizability of the study. Second, this study did not investigate the no-exposure-control Fisher maximum information approach, which many CAT simulation studies use as the baseline, because we tried to focus the comparison on stratified methods instead of other item selection methods. It has been well documented that the no-constraints maximum information approach typically leads to the most accurate ability estimates, but it suffers from highly unbalanced item pool usage, with extremely high exposure rates for some items and low exposure rates for other items. Finally, it may be worthwhile to consider content balancing to make the results more useful for the purpose of practice.

In conclusion, MISTR and UMISTR both produce more precise ability estimation than the traditional *a*-stratified method STR when the test length is long (60 items) and an exposure-control method is used. UMISTR produced slightly less precise ability estimation than USTR but resulted in fewer underused items, indicating a more balanced use of the item pool. MISTR or UMISTR is a viable alternative pool stratification and item selection method when an *a*-stratified procedure is considered for a CAT.

References

Chang, H., Qian, J., & Ying, Z. (2000). *a*-Stratified multistage CAT with *b*-blocking. *Applied Psychological Measurement*, *25*, 333 – 341. https://doi.org/10.1177/01466210122032181

Chang, H., & Ying, Z. (1999). *a*-Stratified computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211 – 222. https://doi.org/10.1177/01466219922031338

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (1999). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing* (Research Report No. ACT-RR-99-5). Iowa City, IA: ACT.

Deng, H., Ansley, T., & Chang, H.-H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, *47*, 202 – 226. https://doi.org/10.1111/j.1745-3984.2010.00109.x

Deng, H., & Chang, H. (2001, April). *a-Stratified computerized adaptive testing with unequal item exposure across strata*. Paper presented at the meeting of the National Council of Measurement in Education, Seattle, WA.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Parshall, C., Davey, T., & Nering, M. (1998, April). *Test development exposure control for adaptive tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Reckase, M. D., & He, W. (2005). *Ideal item pool design for the NCLEX-RN exam*. East Lansing, MI: Michigan State University.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973 – 977). San Diego, CA: Navy Personnel Research and Development Center.

Appendix A: Item Exposure Rate

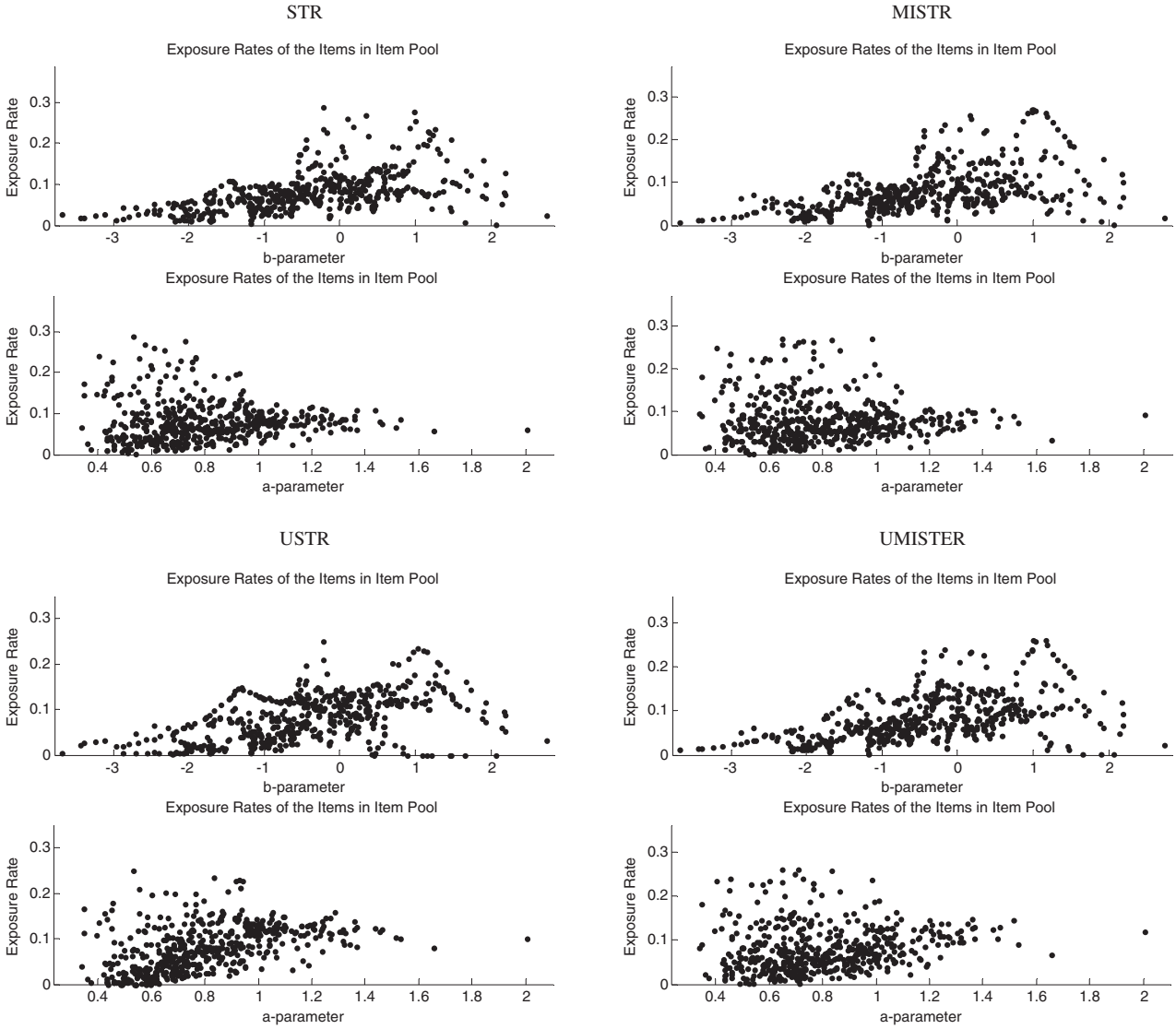

Figure A1 Item exposure rate conditional on items' *a*- or *b*-parameters: 40 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

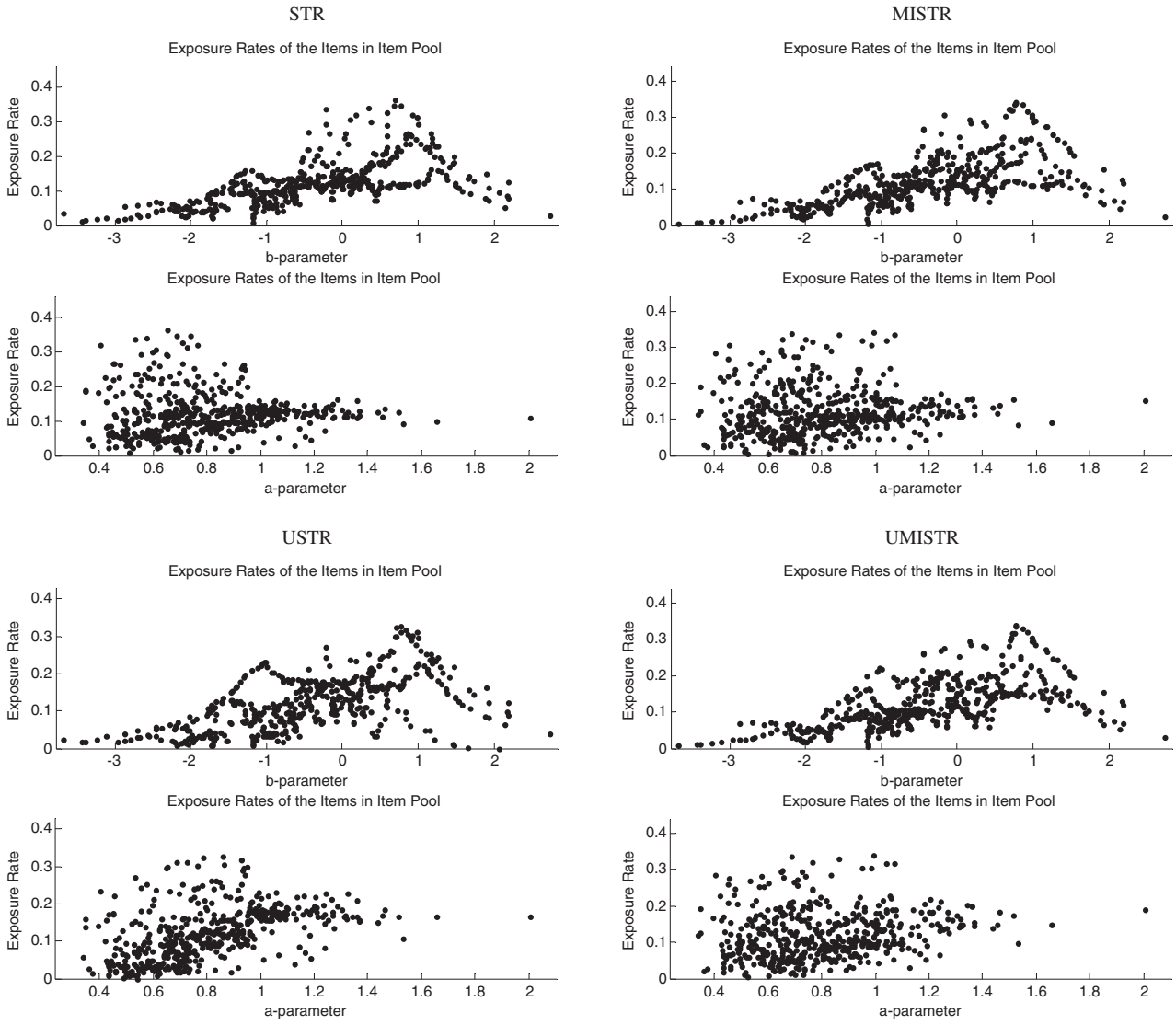

Figure A2 Item exposure rate conditional on items' *a*- or *b*-parameters: 60 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

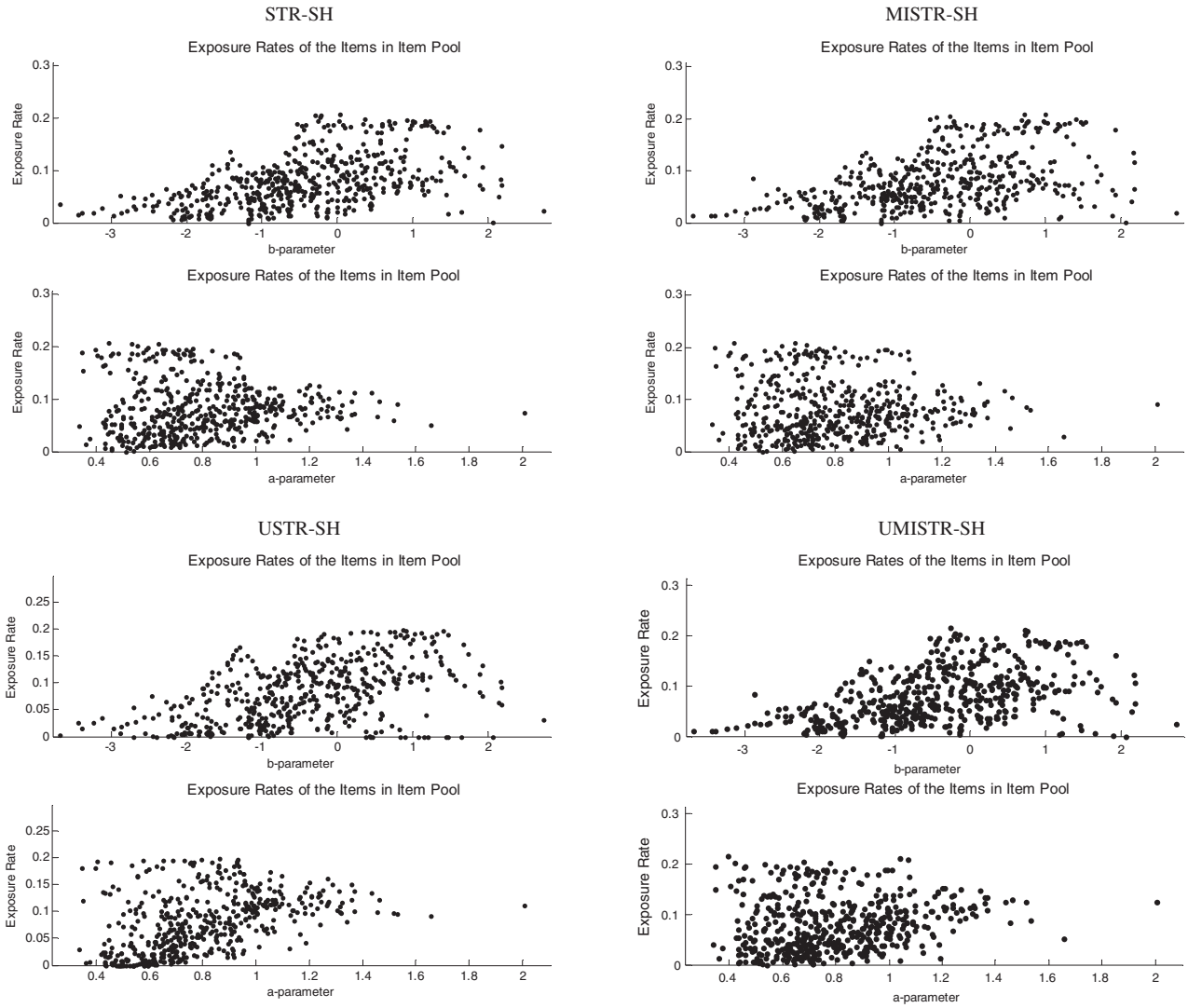

Figure A3 Item exposure rate conditional on item's *a*- or *b*-parameters: 40 items with Sympson – Hetter exposure control. STR-SH = *a*-stratified item selection with Sympson – Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SHl.

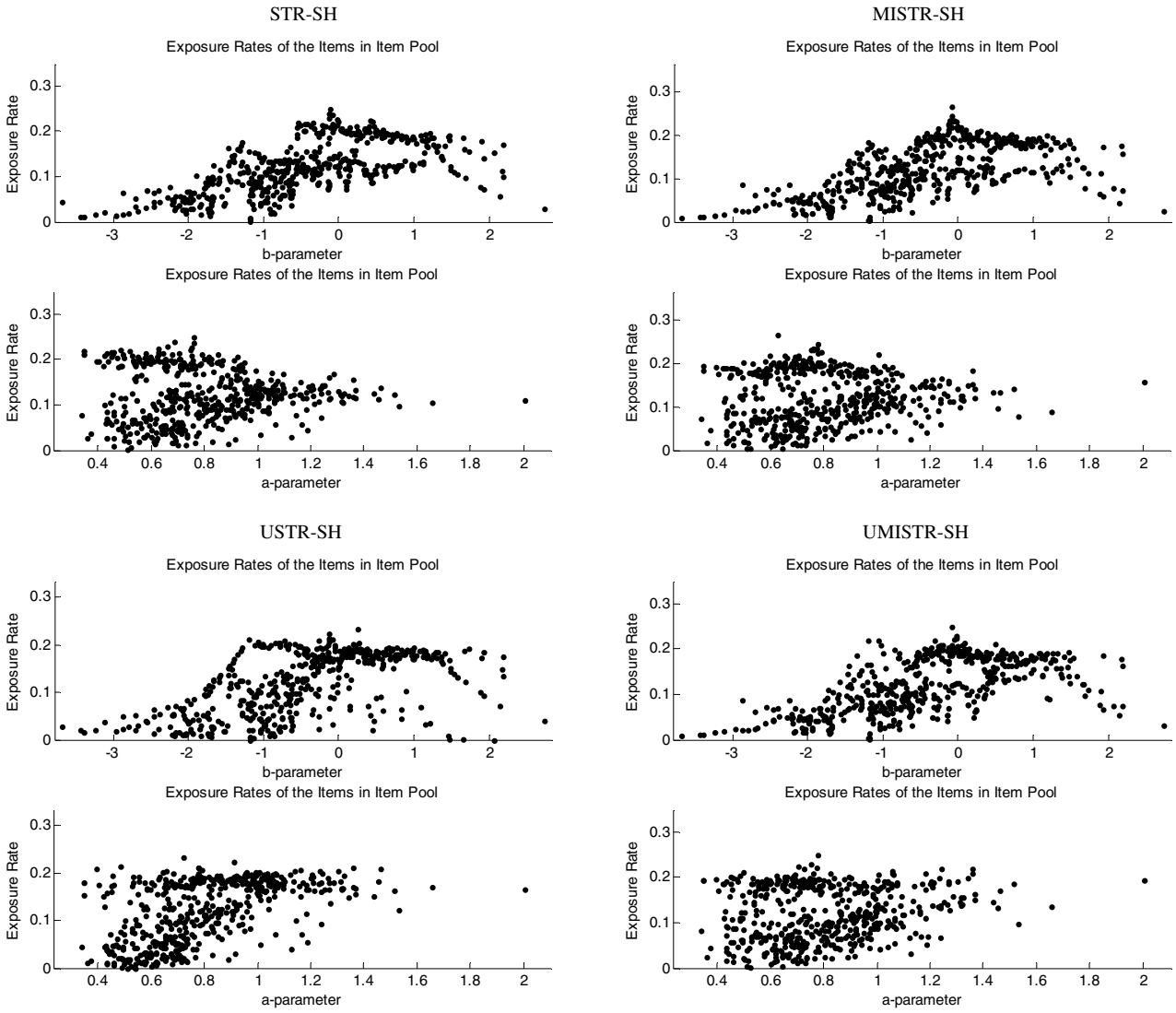

Figure A4 Item exposure rate conditional on item's *a*- or *b*-parameters: 60 items with Sympson – Hetter exposure control. STR-SH = *a*-stratified item selection with Sympson – Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SH.

Appendix B: Box Plots of Ability Estimation

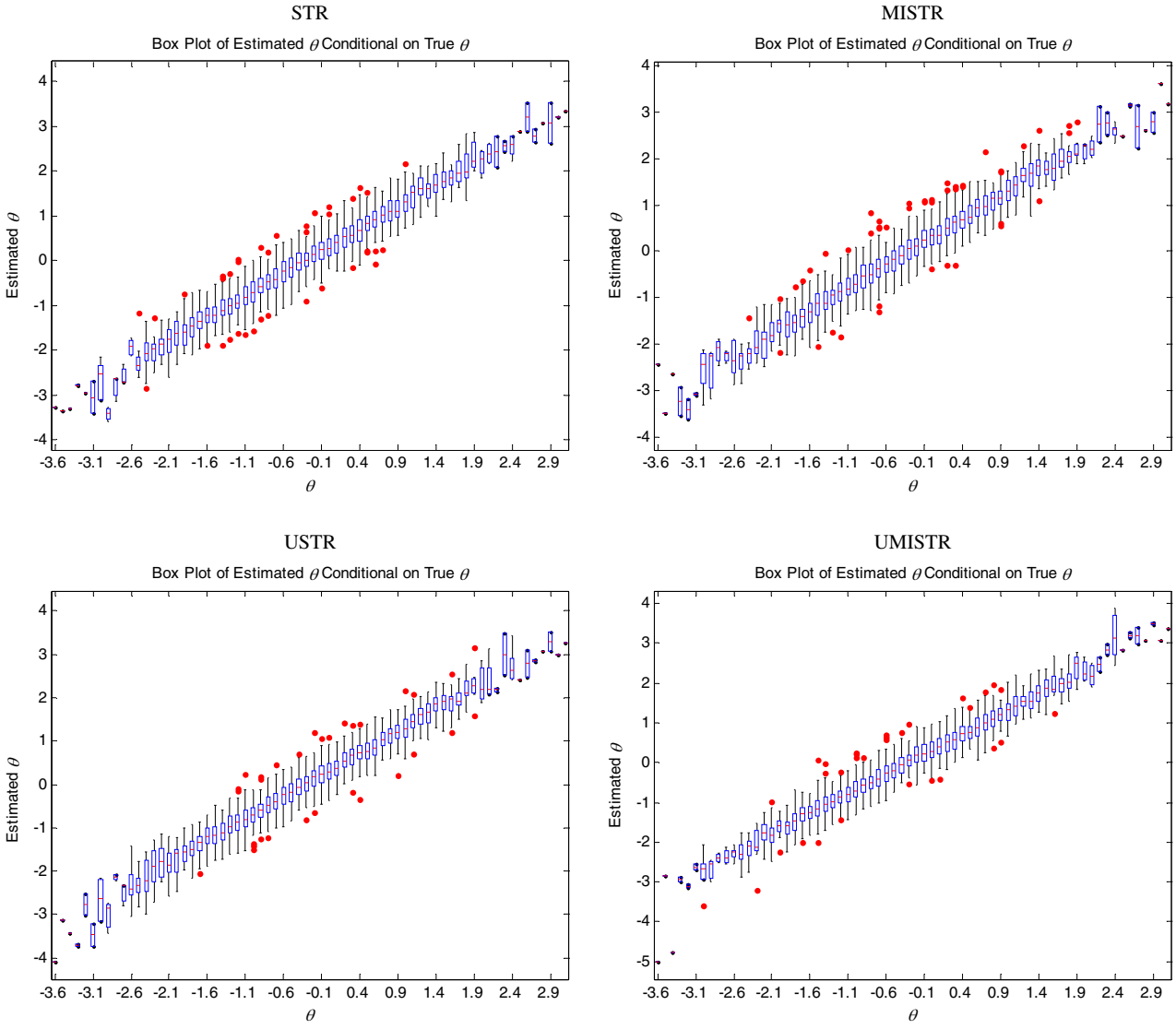

Figure B1 Box plots of ability estimation conditional on true theta: 40 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

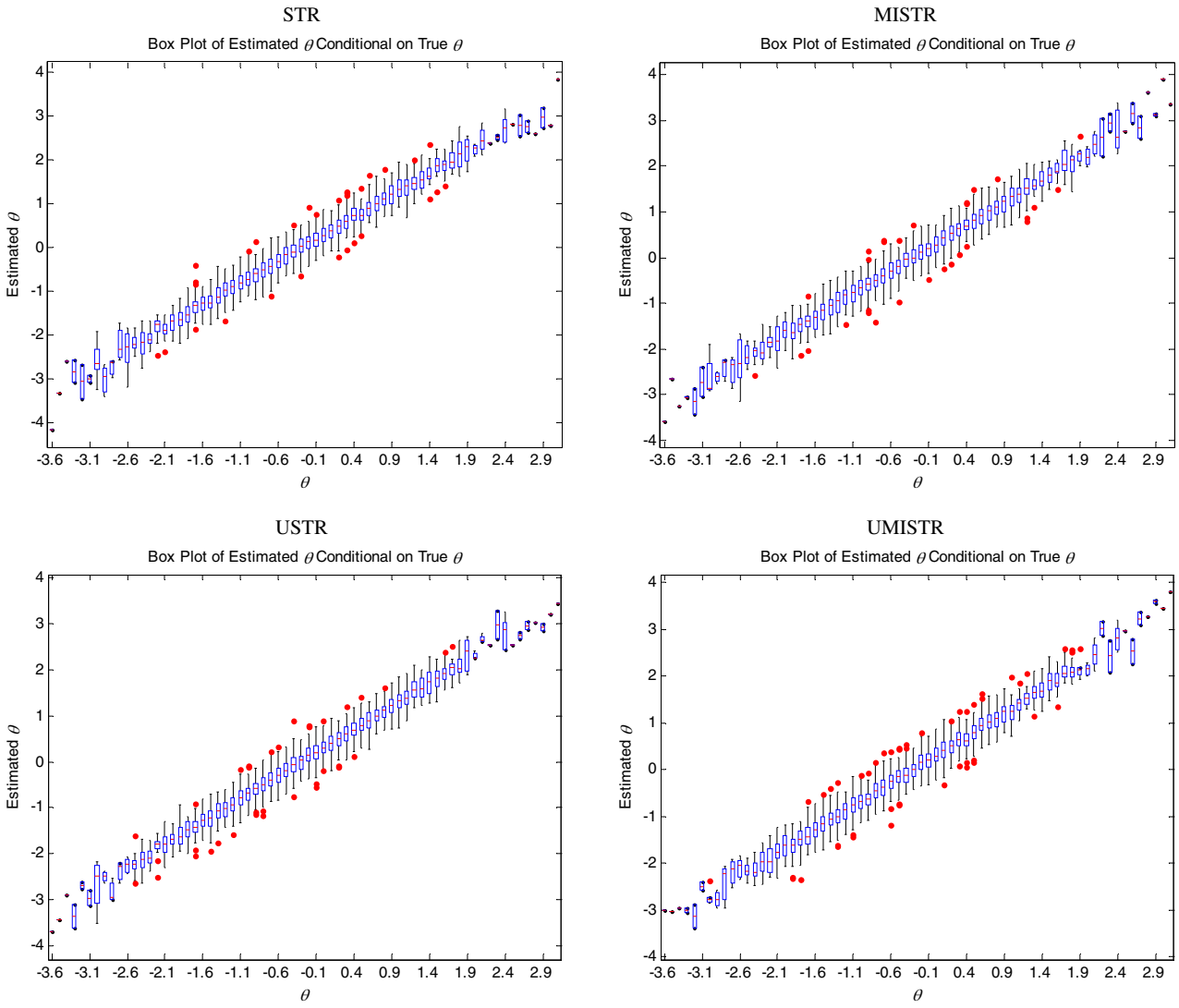

Figure B2 Box plots of ability estimation conditional on true theta: 60 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

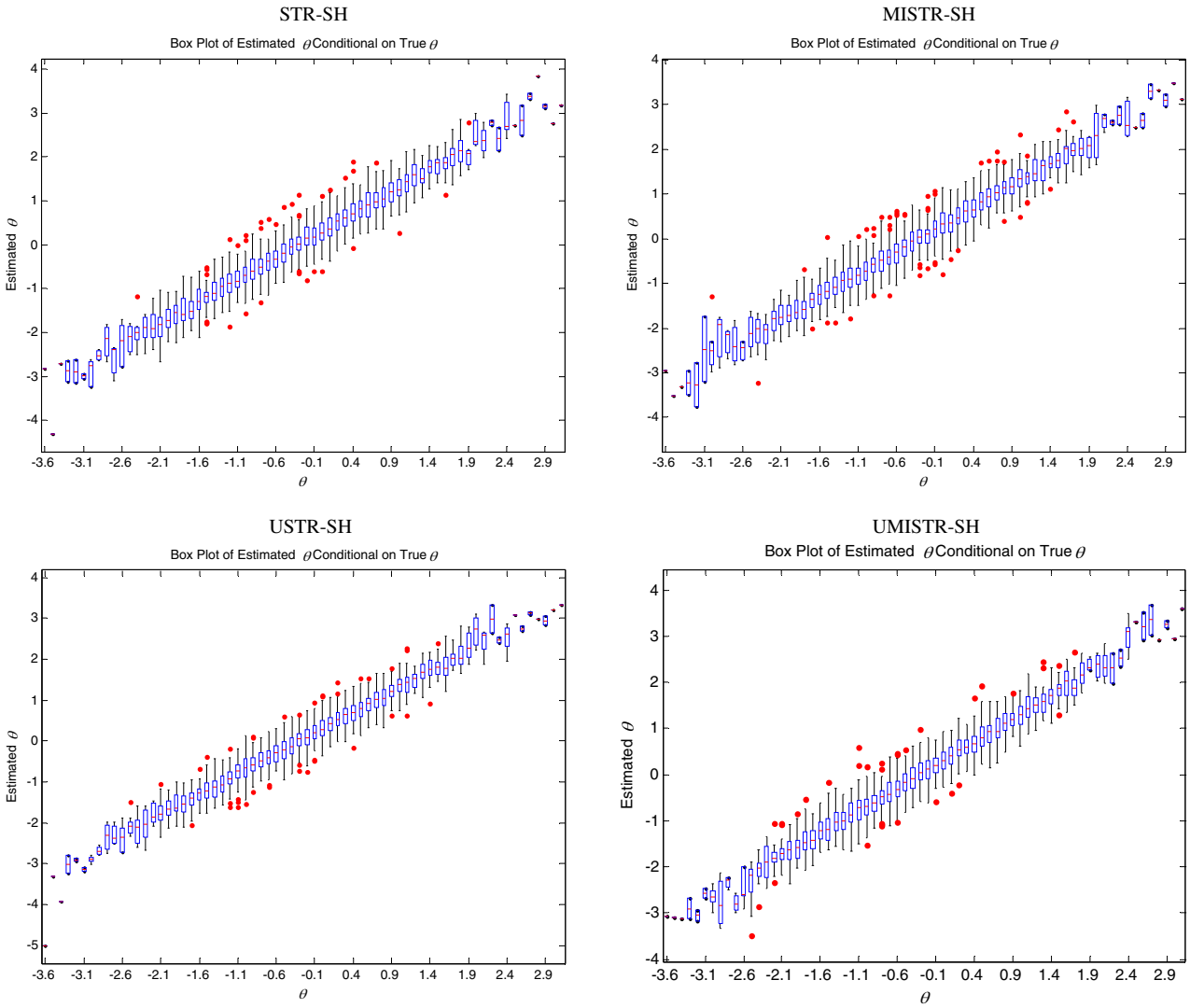

Figure B3 Box plots of ability estimation conditional on true theta: 40 items with Sympson–Hetter exposure control. STR-SH = *a*-stratified item selection with Sympson–Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SH.

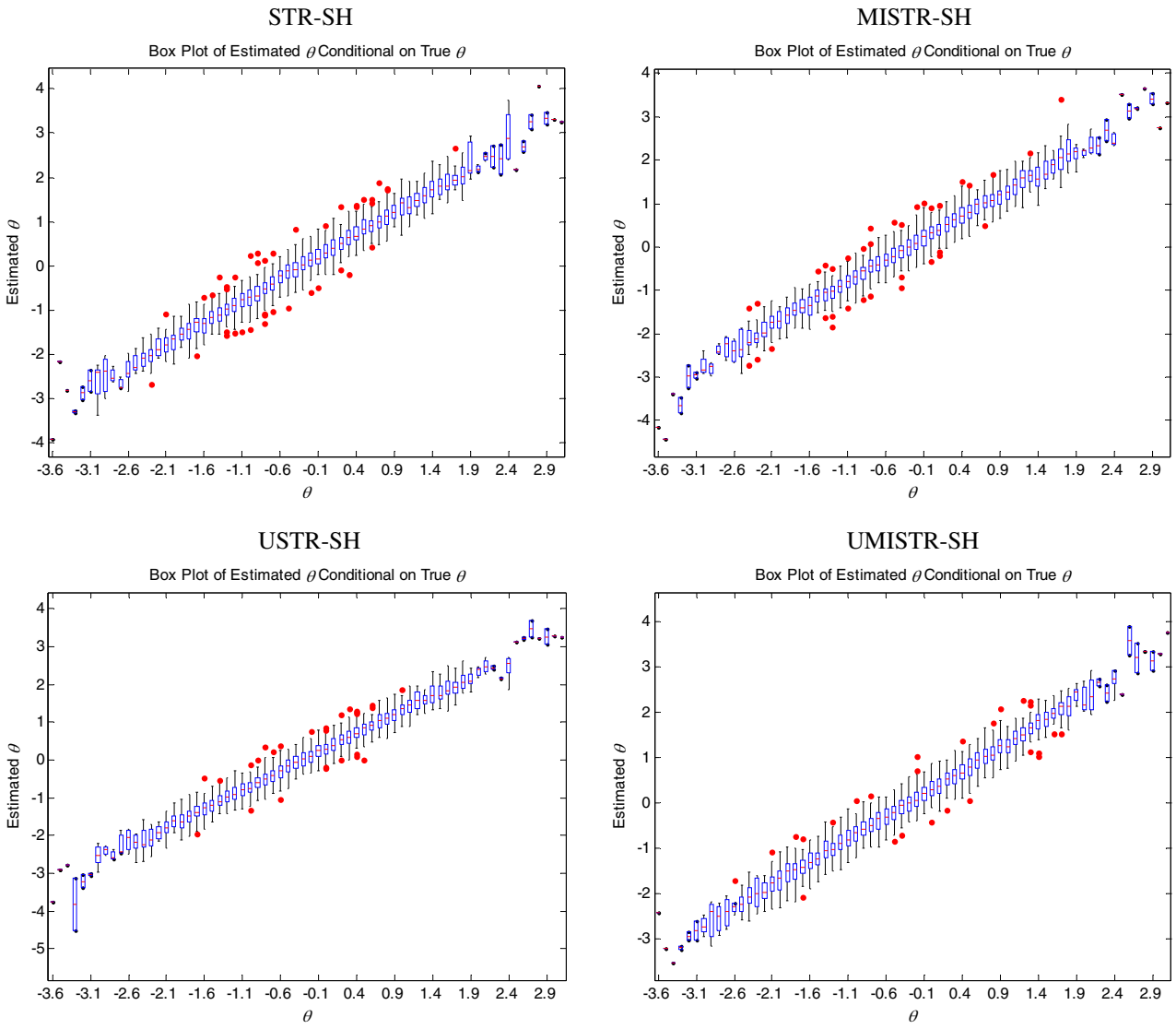

Figure B4 Box plots of ability estimation conditional on true theta: 60 items with Sympson–Hetter exposure control. STR-SH = *a*-stratified item selection with Sympson–Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SH.

Appendix C: Test Information Function

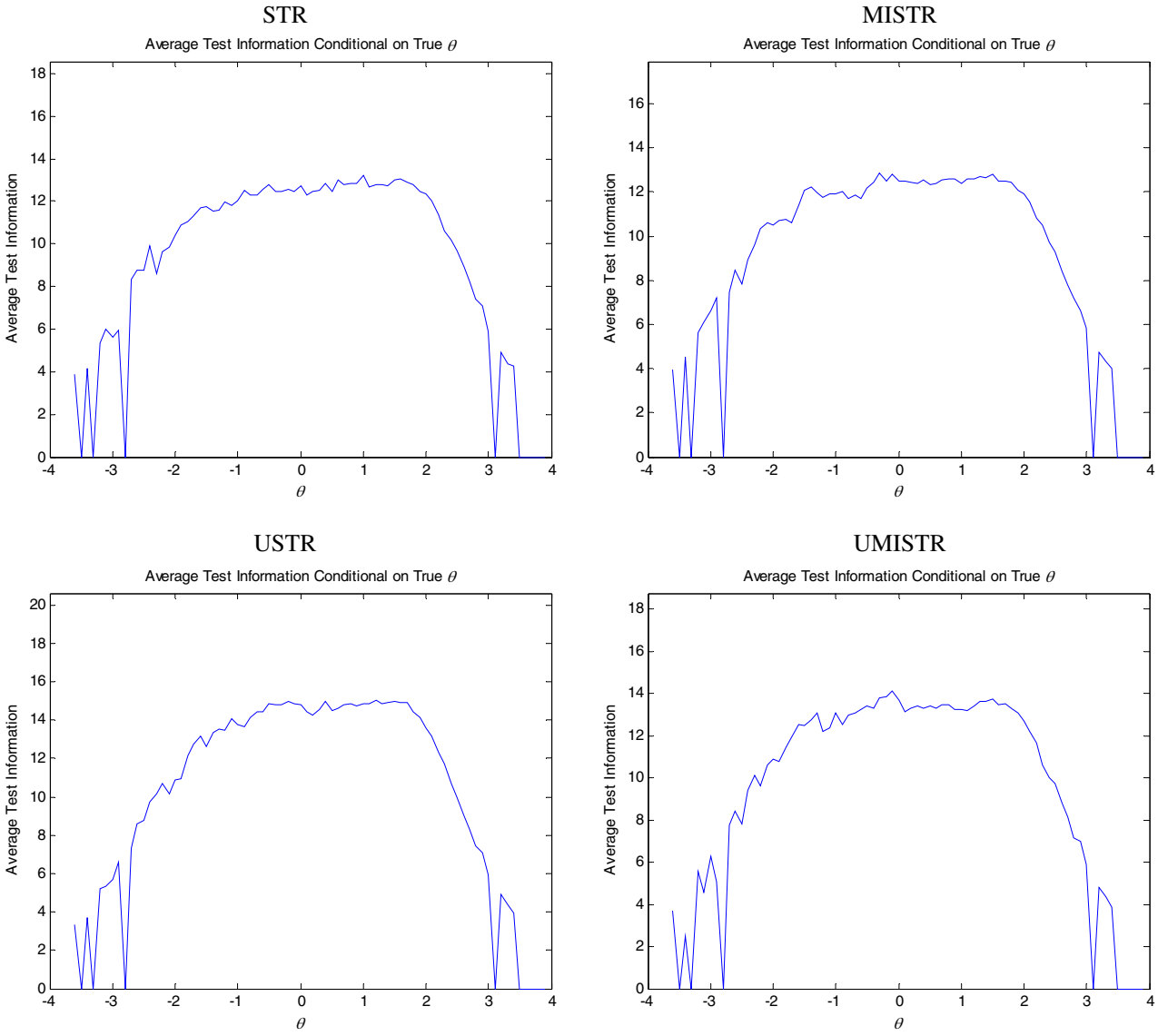

Figure C1 Average test information function conditional on true theta: 40 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

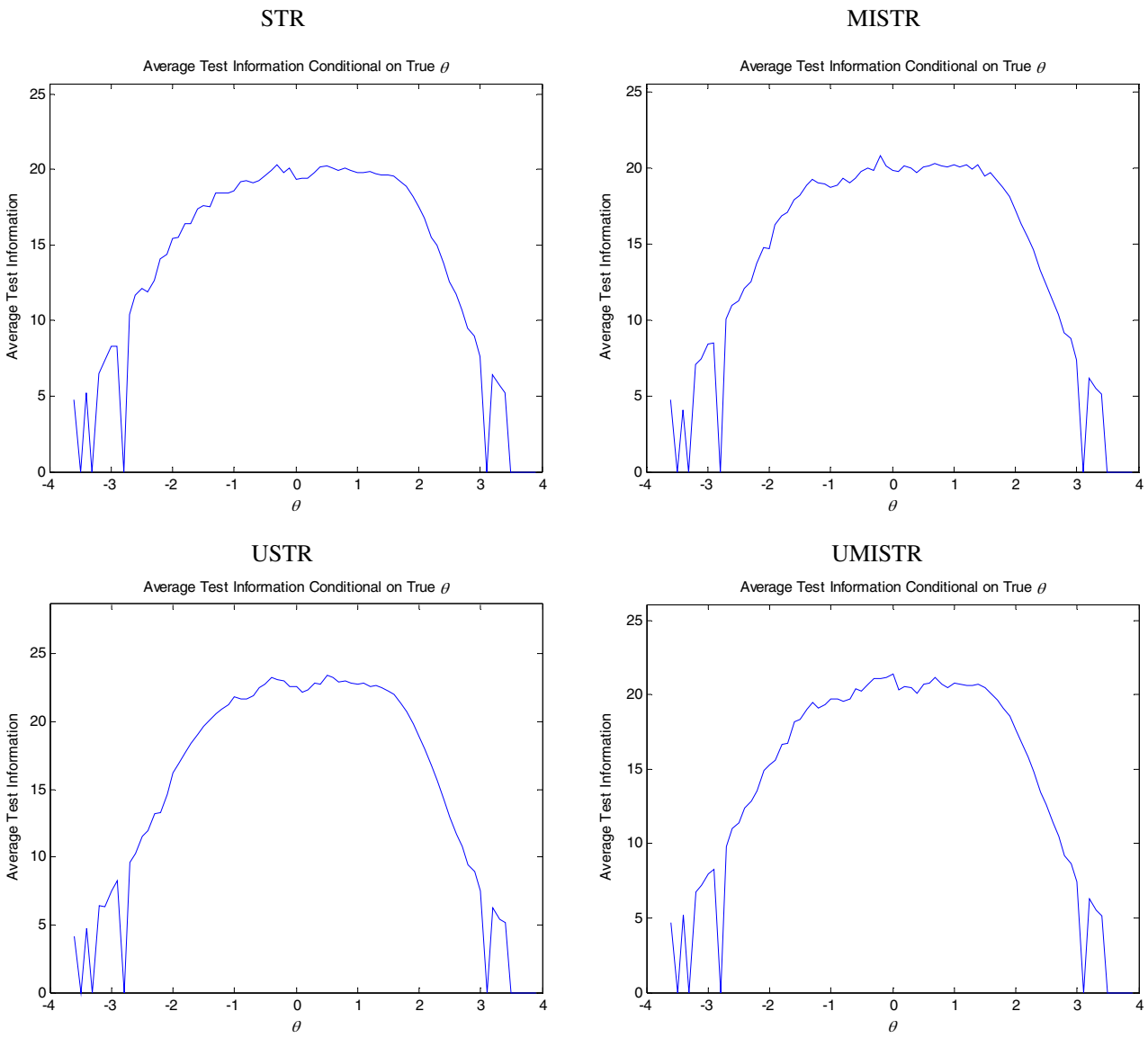

Figure C2 Average test information function conditional on true theta: 60 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

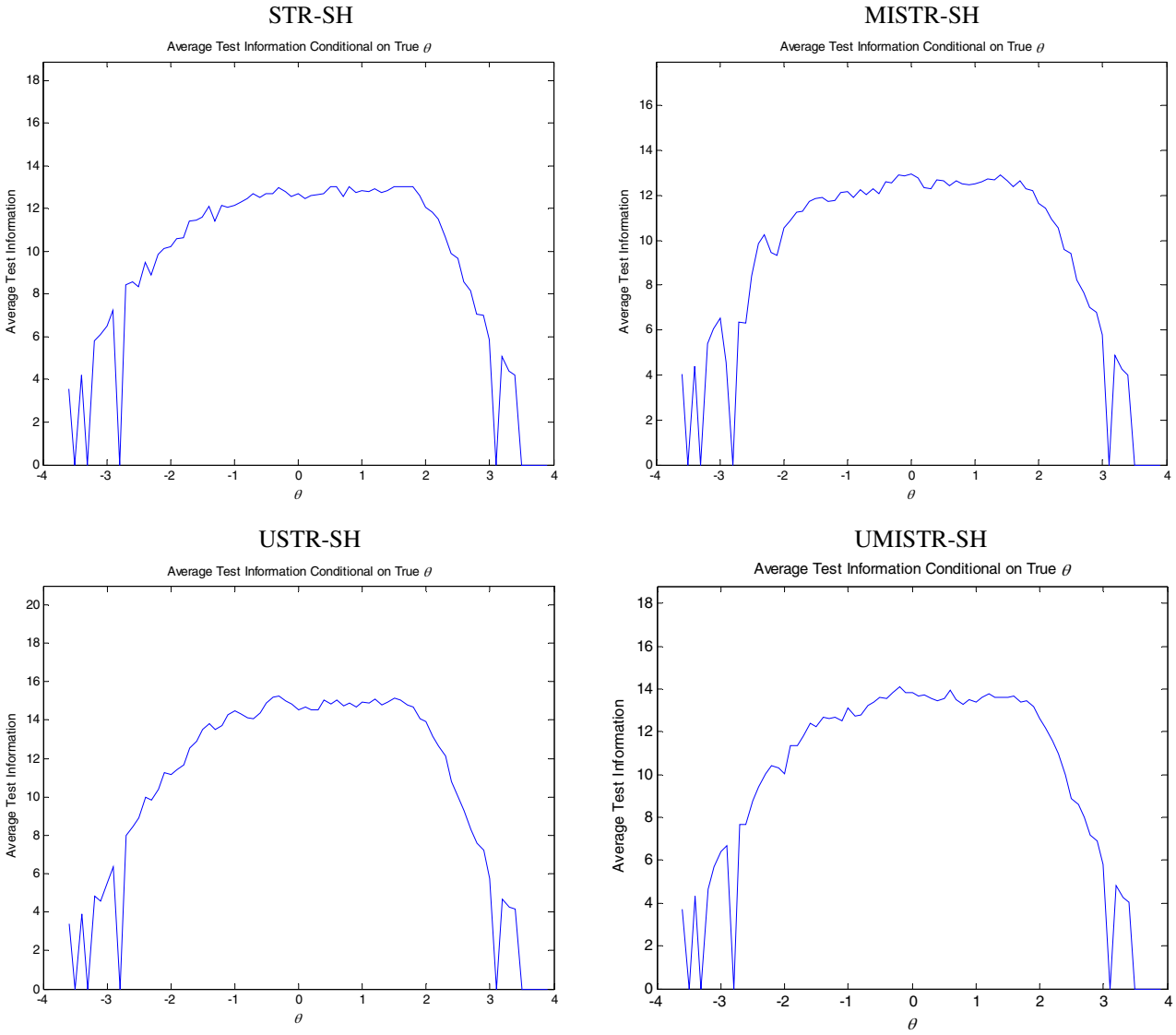

Figure C3 Average test information function conditional on true theta: 40 items with Sympson – Hetter exposure control. STR-SH = *a*-stratified item selectionwith Sympson – Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SH.

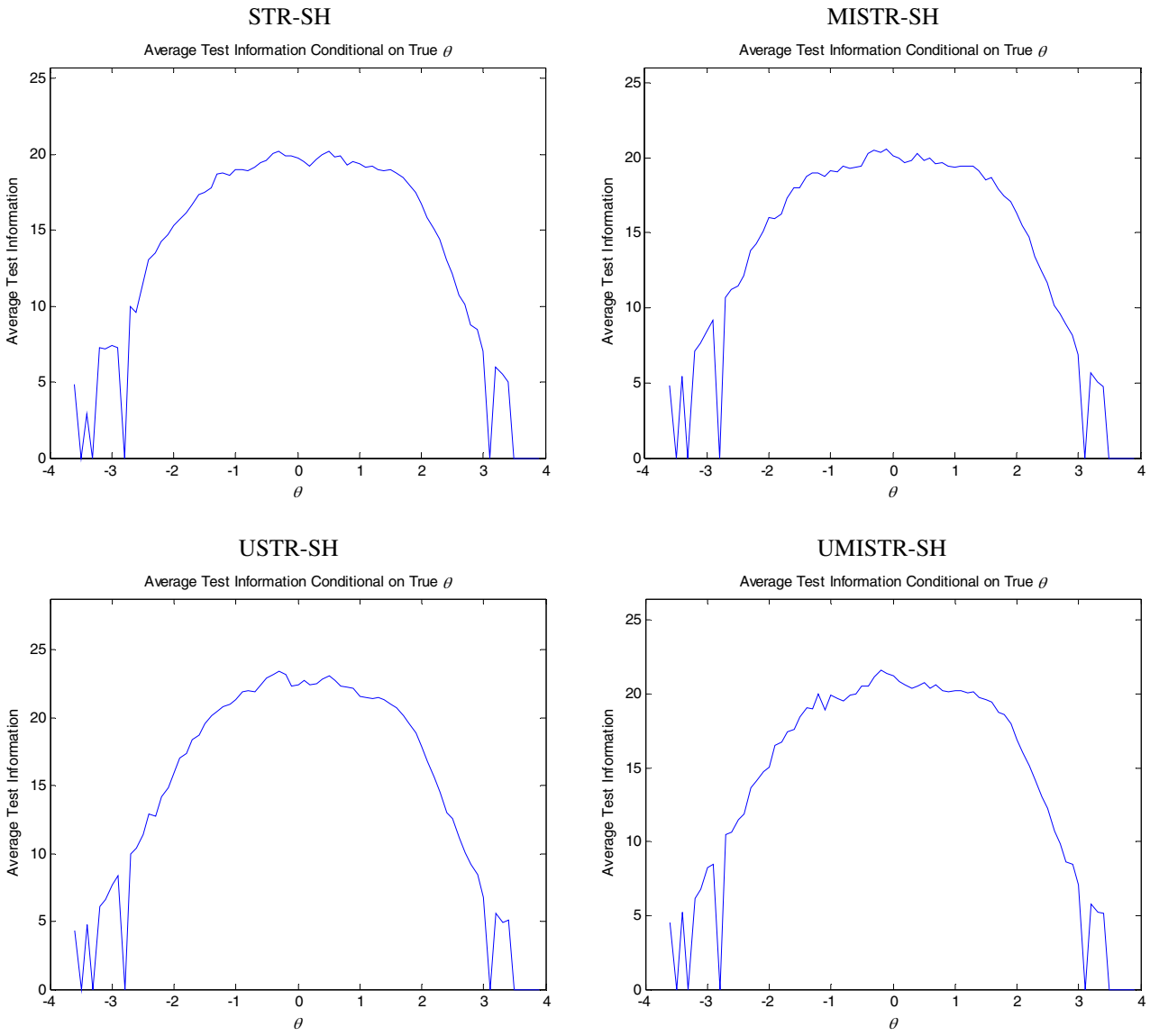

Figure C4 Box plot of ability estimation conditional on true theta: 60 items with Sympson–Hetter exposure control. STR-SH = *a*-stratified item selectionwith Sympson–Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SH.

Appendix D: Average Conditional Standard Errors of Measurement

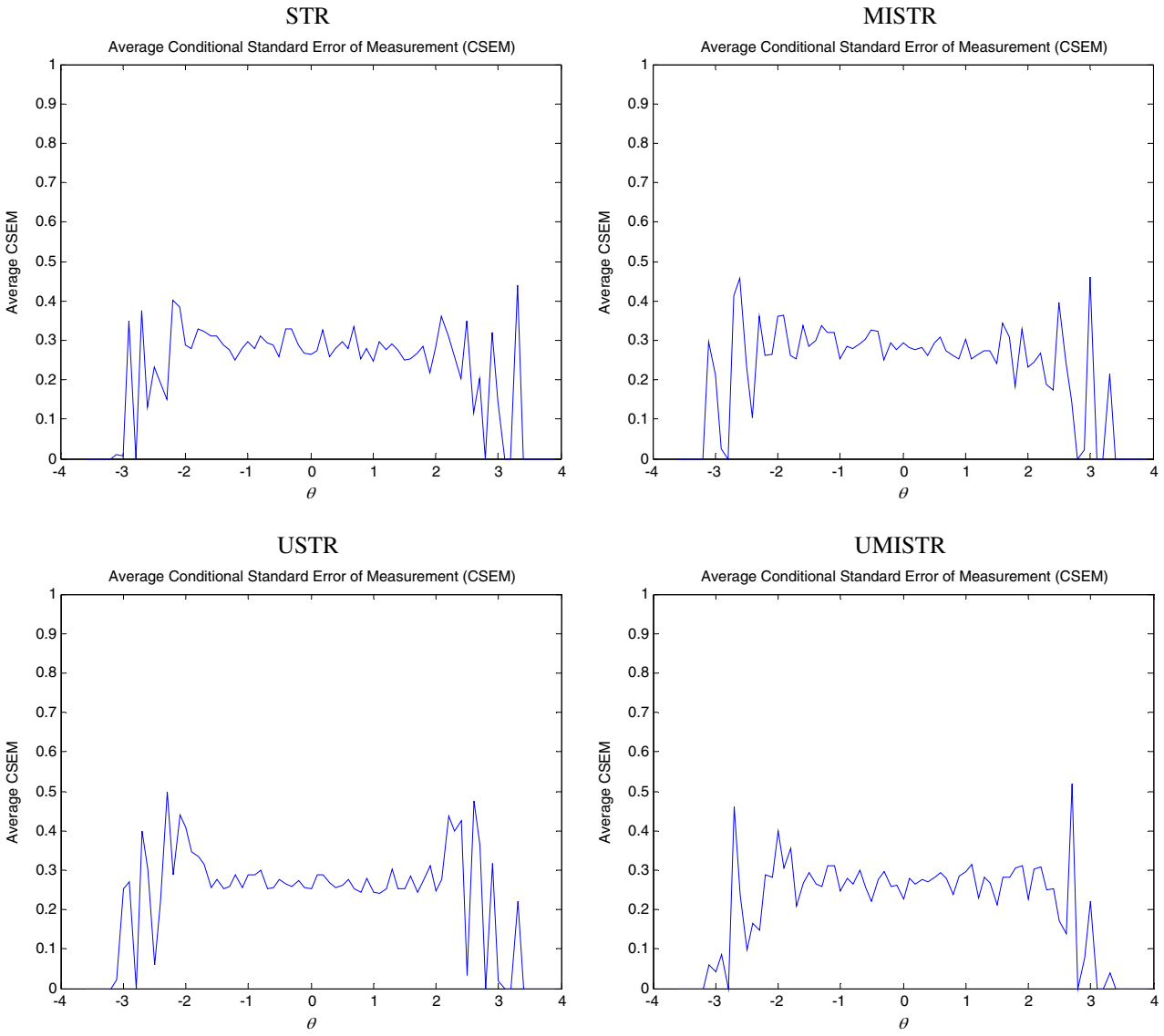

Figure D1 Average standard errors of measurement conditional on true theta: 40 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

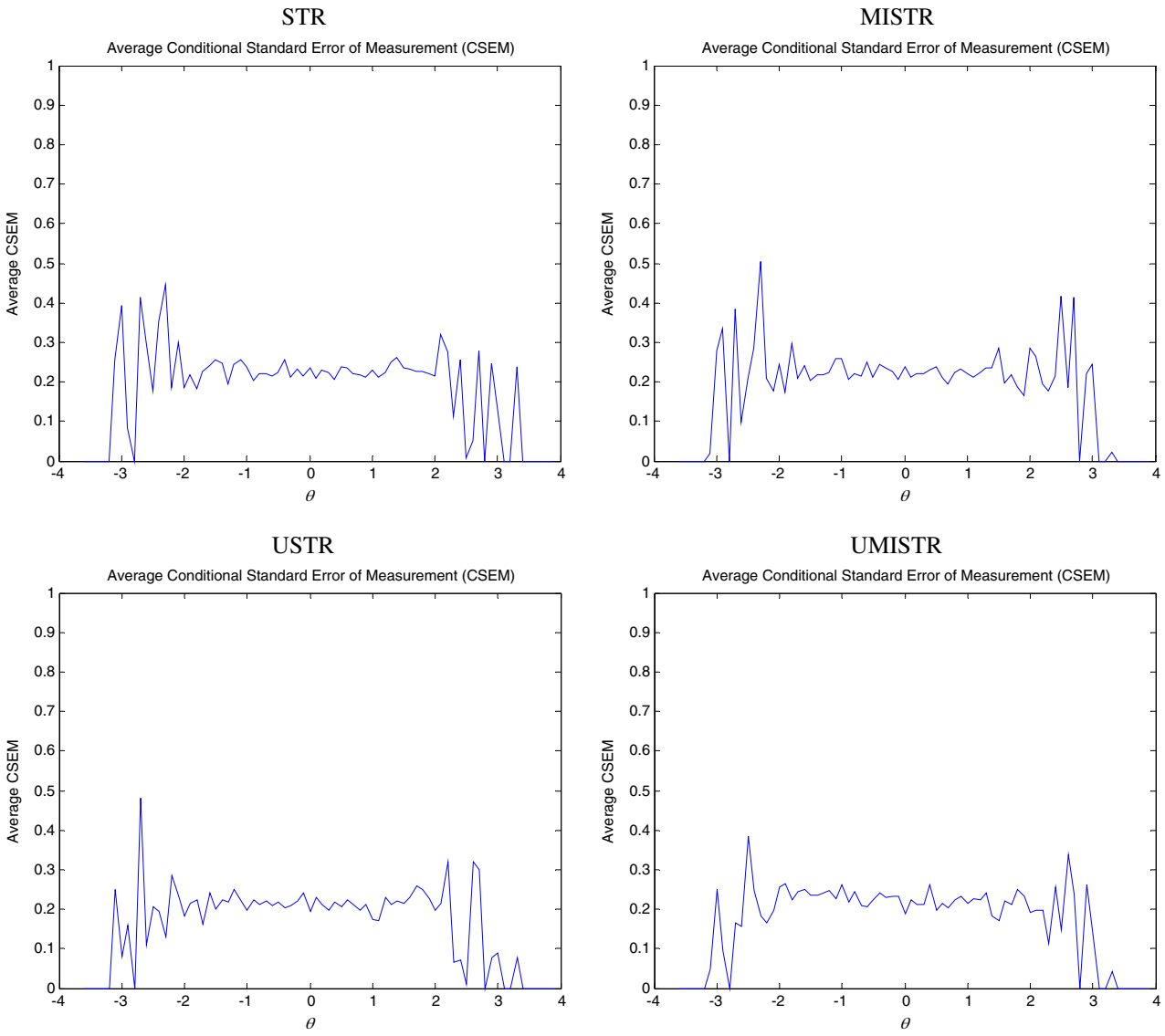

Figure D2 Average standard errors of measurement conditional on true theta: 60 items without exposure control. STR = *a*-stratified item selection; USTR = unequal STR; MISTR = maximum information STR; UMISTR = unequal maximum information STR.

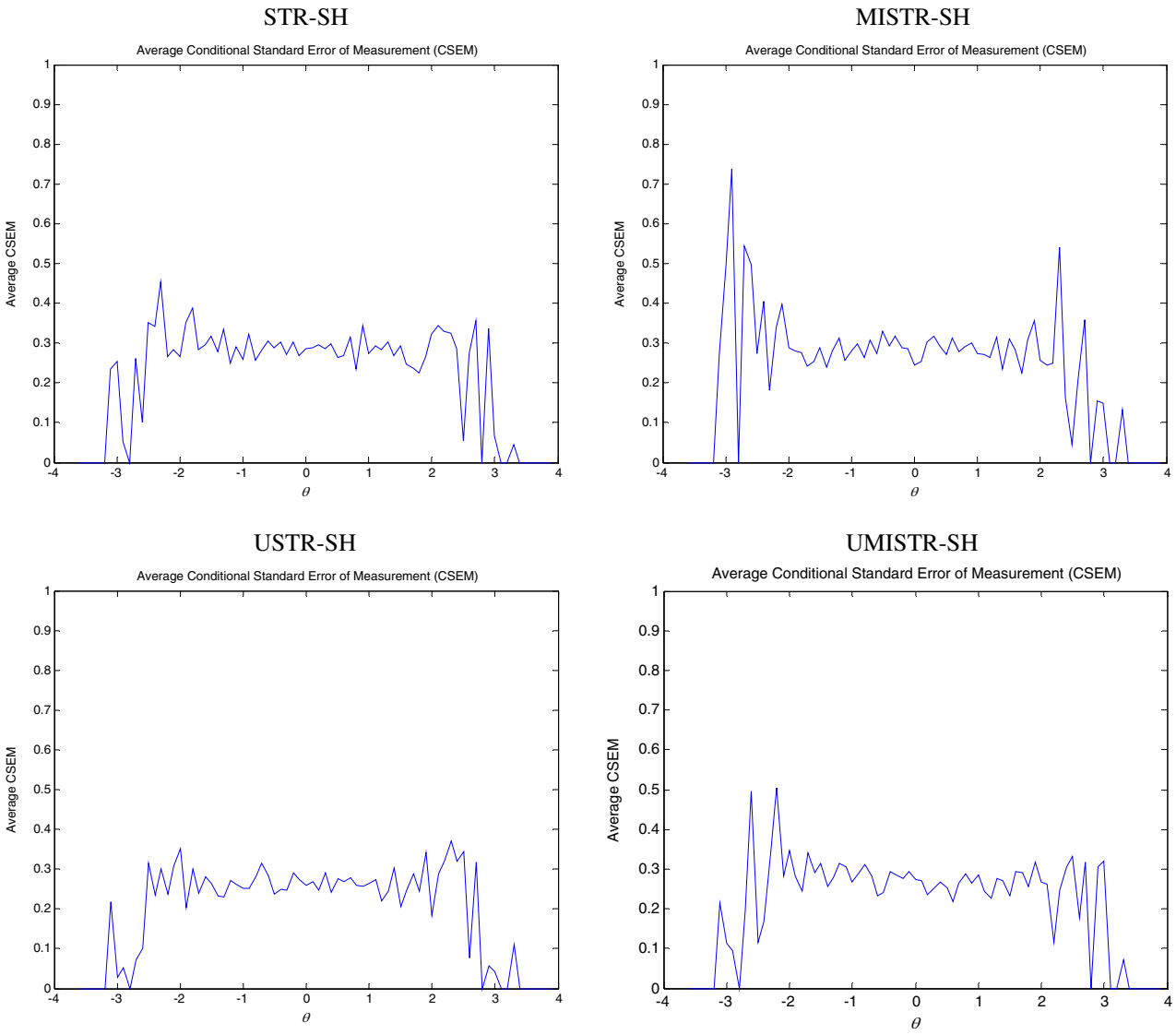

Figure D3 Average standard error of measurement conditional on true theta: 40 items with Sympson–Hetter exposure control. STR-SH = *a*-stratified item selectionwith Sympson–Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SH.

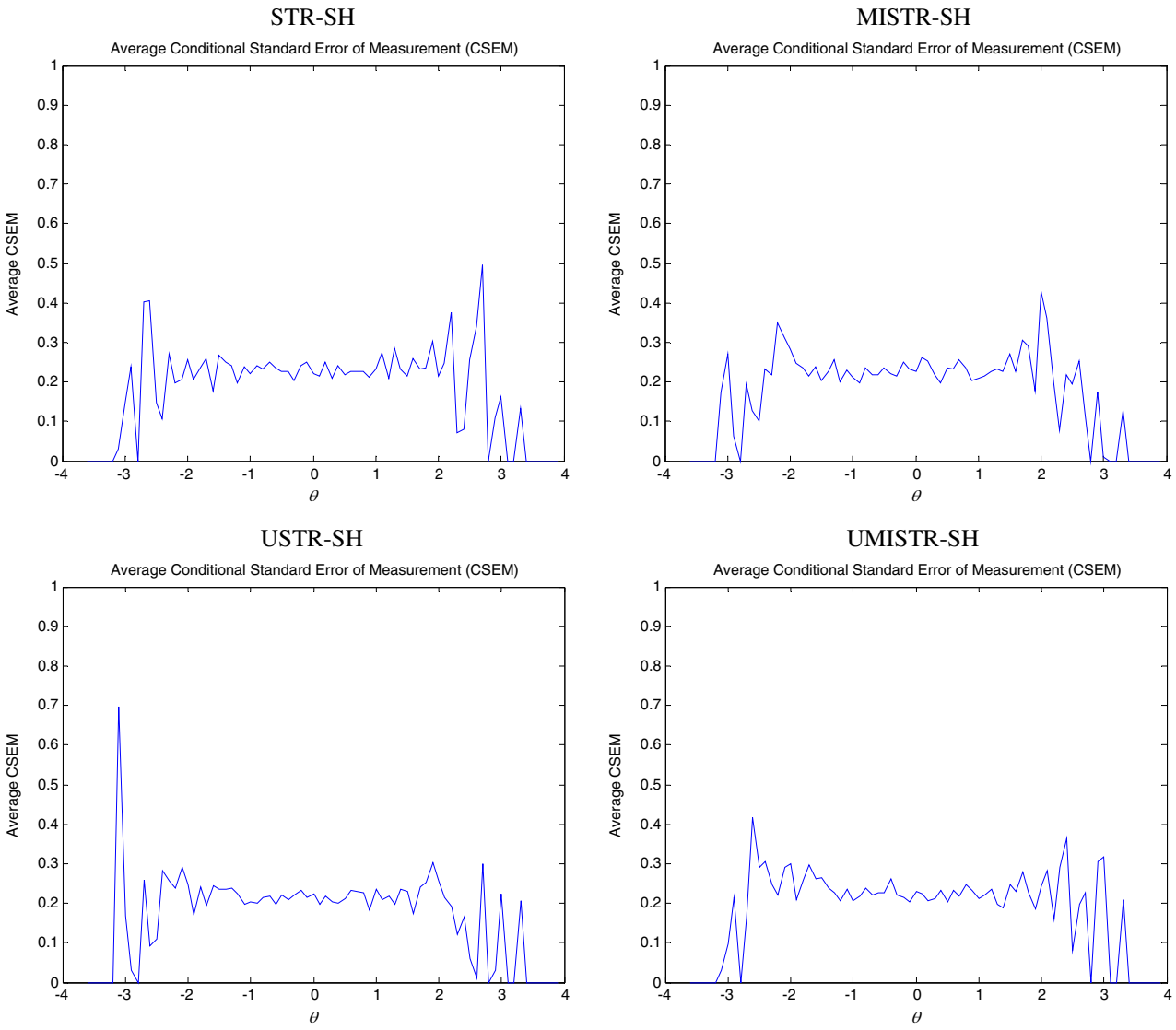

Figure D4 Average standard errors of measurement conditional on true theta: 60 items with Sympson–Hetter exposure control. STR-SH = *a*-stratified item selection with Sympson–Hetter exposure control; USTR-SH = unequal STR-SH; MISTR-SH = maximum information STR-SH; UMISTR-SH = unequal maximum information STR-SH.

Suggested citation:

Gu, L., Ling, G., & Qu, Y. (2019). *A modified a-stratified method for computerized adaptive testing* (Research Report No. RR-19-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12246

Action Editor: Rebecca Zwick

Reviewers: Qiwei He and Duanli Yan